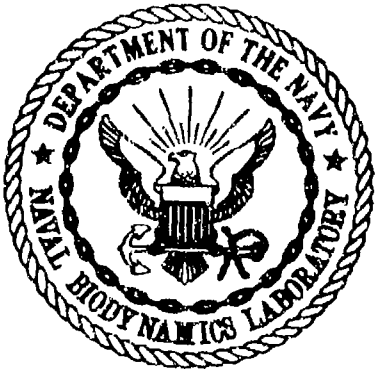
NBDL-M001

LEVEL II

(1)

STABILIZATION AND TASK DEFINITION IN A PERFORMANCE TEST BATTERY

Marshall B. Jones



October 1980

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NBDL M001	2. GOVT ACCESSION NO.: AD-A099987
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Stabilization and Task Definition in a Performance Test Battery	5. TYPE OF REPORT & PERIOD COVERED: Monograph, Period Ending May 12, 1979
6. PERFORMING ORG. REPORT NUMBER:	
7. AUTHOR(s): Marshall B. Jones	8. CONTRACT OR GRANT NUMBER(s): Contract No. N0023-79-M-5089
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Behavioral Sciences, Milton S. Hershey Medical Center, Penn St. Univ., Hershey, PA 17033	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Project No. F58524, Task Area ZF5852406, Work Unit MF58 524 002 5027
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Medical Research & Development Command, Bethesda, MD 20014	12. REPORT DATE: May 12, 1979
	13. NUMBER OF PAGES: 59
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Naval Biodynamics Laboratory, Box 29407, New Orleans, LA 70189	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Final Report on Navy Contract No. N0023-79-M-5089

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Performance, Skill Acquisition, Task Stabilization, Individual Differences, Performance Testing, Environmental Stress

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Performance testing under unusual environmental circumstances almost always involves repeated-measure designs. Most tasks, however, show practice effects with repeated administrations, effects that may appear in the group mean, the variance among subjects, or the correlations over subjects among trials or repeated testings. There comes a point in many tasks after which practice no longer produces changes in performance; as we will put it, the task stabilizes. Our criteria for stabilization are: the group mean no longer increases, or increases at a slow and regular rate; the variance among subjects no longer

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

changes; and the correlation with earlier trials remains the same from one stabilized trial to the next; finally, the correlation among stabilized trials is constant. Stabilization in this sense is virtually essential to a performance test battery. When it is absent, practice and environmental effects are confounded; interpretation becomes very difficult; and problems of design are greatly complicated, in some cases impossibly so. It is desirable, therefore, to determine in advance whether or not a task stabilizes with practice and, if so, how long it takes. It is additionally desirable that a task be well defined, that is, that it stabilize at a high level; preferably the average correlation among stabilized trials should be greater than .90. The present report concerns ten tasks each of which was practiced for 15 days by either 18 or 19 subjects. The ten tasks were: Complex Counting, Grammatical Reasoning, Code Substitution, Stroop Color-Words, Arithmetic, Letter Search, Critical Tracking, Compensatory Tracking, Time Estimation, and Spoke Trail Making. The ten tasks were practiced in the order given. All subjects were Navy enlisted men between 19 and 24 years old and with 20/20 vision. The ten tasks were all analyzed according to the criteria mentioned above, beginning with the mean and variance and then determining the stability of cross session reliabilities. Analysis of the ten tasks was straightforward and according to the criteria mentioned above with respect to the means and standard deviations. In order to determine the stability of the correlation among trials, a series of two way ANOVA's is applied to the correlation matrices. Each of the ten tasks was subjected to this same step-wise analysis. Some tasks, for example, Arithmetic, stabilized completely. Others stabilized in some respects but not in others. Some tasks had more than one dependent measure and in these cases stabilization sometimes occurred in one dependent measure when it did not occur in another. The bulk of the report is given over to detailing these results and describing the application of the ANOVA employed.

NBDL - M001

STABILIZATION AND TASK DEFINITION IN A PERFORMANCE TEST BATTERY

Marshall B. Jones

October 1980

Bureau of Medicine and Surgery
Work Unit MF58.524.002-5027
Navy Contract N0023-79-M-5089

Approved by

Channing L. Ewing, M. D.
Scientific Director

Released by

Captain J. E. Wenger MC USN
Commanding Officer

Naval Biodynamics Laboratory
Box 29407
New Orleans, LA 70189

ABSTRACT

Performance testing under unusual environmental circumstances almost always involves repeated-measure designs. Most tasks, however, show practice effects with repeated administrations, effects that may appear in the group mean, the variance among subjects, or the correlations over subjects among trials or repeated testings. There comes a point in many tasks after which practice no longer produces changes in performance; as we will put it, the task stabilizes. Our criteria for stabilization are: the group mean no longer increases, or increases at a slow and regular rate; the variance among subjects no longer changes; and the correlation with earlier trials remains the same from one stabilized trial to the next; finally, the correlation among stabilized trials is constant. Stabilization in this sense is virtually essential to a performance test battery. When it is absent, practice and environmental effects are confounded; interpretations becomes very difficult; and problems of design are greatly complicated, in some cases impossibly so. It is desirable, therefore, to determine in advance whether or not a task stabilizes with practice and, if so, how long it takes. It is additionally desirable that a task be well defined, that is, that it stabilize at a high level; preferably the average correlation among stabilized trials should be greater than .90. The present report concerns ten tasks each of which was practiced for 15 days by either 18 or 19 subjects. The ten tasks were: Complex Counting, Grammatical Reasoning, Code Substitution, Stroop Color-Words, Arithmetic, Letter Search, Critical Tracking, Compensatory Tracking, Time Estimation, and Spoke Trail Making. The ten tasks were practiced in the order given. All subjects were Navy enlisted men between 19 and 24 years old and with 20/20 vision. The ten tasks were all analyzed according to the criteria mentioned above, beginning with the mean and variance and then determining the stability of cross session reliabilities. Analysis of the ten tasks was straightforward and according to the criteria mentioned above

with respect to the means and standard deviations. In order to determine the stability of the correlation among trials, a series of two way ANOVA's is applied to the correlation matrices. Each of the ten tasks was subjected to this same step-wise analysis. Some tasks, for example, arithmetic, stabilized completely. Others in some respects but not in others. Some tasks had more than one dependent measure and in these cases, stabilization sometimes occurred in one dependent measure when it did not occur in another. The bulk of the report is given over to detailing these results and describing the application of the ANOVA employed.

Marshall B. Jones is with the Department of Behavioral Sciences, Milton S. Hershey Medical Center, Pennsylvania State University, Hershey, PA 17033.

Final Report on Navy Contract No. N0023-79-M-5089, May 12, 1979.

STABILIZATION AND TASK DEFINITION IN A PERFORMANCE TEST BATTERY

Marshall B. Jones

Kennedy and Bittner (1977) have recently detailed the need for a performance test battery to study the effects of unusual environments over prolonged exposure periods. The same authors also point out that performance testing in environmental research almost always involves repeated-measure designs. This latter circumstance has definite consequences for the properties that a performance test or battery should have. When a task or test is administered on repeated occasions, it usually shows practice effects; and these effects may appear in the mean, the variance among subjects, or in the correlations over subjects among trials or repetitions. If practice is continued, there comes a point in many tasks after which practice effects no longer appear; as we will put it, the task stabilizes. The mean becomes asymptotic or increases at a slow and regular rate; the variance among subjects remains the same from trial to trial; and the correlation with trials earlier in the practice sequence remains the same from one stabilized trial to another; in addition, the correlation between any two stabilized trials is constant. Not all tasks stabilize, however, and among those that d some stabilize more quickly than others (Jones, 1972). Furthermore, different tasks may stabilize at different levels; that is, the average correlation among stabilized trials may vary from one task to another (Jones, 1970 a & b).

If a task does not or has not been stabilized in a group of subjects, its use in a repeated-measure design is compromised from the start. If the data are analyzed by univariate analysis of variance, one of the requirements

of a repeated-measure design, compound symmetry, may not be met, with serious consequent difficulties for the analysis (Winer, 1971). These difficulties can be overcome in large measure by resort to multivariate statistical methods, but only if the subjects outnumber the repeated measurements, preferably by a considerable margin (Morrison, 1967). This condition, however, is often difficult or impossible to meet in field experiments under unusual environmental circumstances.

If the battery or tests from it are used to monitor individual performance, further difficulties arise. If a task has not been stabilized, the correlations among successive trials will very likely show "superdiagonal form" (Jones, 1969a). That is, the correlation between two trials decreases with the separation between them and, hence, is largest when one trial immediately follows the other. This pattern has been interpreted by some workers to mean that the differential composition of the task is changing and by others to mean that the abilities possessed by the subjects are changing (Alvares and Hulin, 1972, 1973; Dunham, 1974). Under either interpretation an individual's performance could deteriorate or improve over a given span of testings for reasons that have nothing to do with concurrent environmental stresses or events. If the task is changing, the subject may do poorly because he happens to be weak in the abilities or other factors that are prominent in the differential composition of the test over that particular span of testings. If the subject's abilities are changing, then clearly his or her performance may change also and altogether independently of external factors.

The presence of superdiagonal form also makes it difficult to know "what is being measured." To begin with, there is the ambiguity as to basic inter-

pretation. Is it the task or the subject who is changing or, perhaps, both? If the task is changing, then the interpretation of performance changes with every stage of practice. If the subject is changing, then he or she possesses a somewhat different mix of relevant abilities at every stage of practice.

For all these reasons a test, to be included in a performance test battery for environmental research, should stabilize. For many trials or administrations the test may show practice effects but there must come a point after which the test (or subject) no longer changes. It is additionally desirable that the test stabilize at a high level, preferably greater than .80. We will call the level at which a task stabilizes, that is, the average correlation among stabilized trials, task definition. A task is well or poorly defined according as this average ranges up or down from .80.

The primary concern of the Naval Aerospace Medical Research Laboratory Det. is with inertial environments, a particular interest being the very low frequency motions ($<$ 1 Hz) which occasion seasickness and air sickness. In this connection, a research program is underway to develop a test battery for evaluating the performance of a subject who may be exposed to such motions. The general plan of this Performance Evaluation Test for Environmental Research (PETER) is discussed elsewhere (Kennedy and Bittner, 1977); other findings are reported in Kennedy and Bittner (1978 a & b).

This report concerns ten tasks each of which has been practiced for 15 days by 18 or 19 volunteer subjects. In each case our chief concern will be whether or not the task stabilizes and, if so, after how many trials. The remainder of the report is organized into three sections. The first develops the analysis to be used throughout the report, with Critical Tracking serving as an illustrative task.

The second section presents the findings for five tasks all of which stabilize quickly and have acceptable task definition: Code Responses, Grammatical Reasoning, Arithmetic, Stroop Color-Words, and Two-Dimensional Tracking. (Critical Tracking also stabilizes with acceptable task definition.)

The third section presents results concerning four tasks (Complex Counting, Time Estimation, Letter Search, and the Spoke Trail-Making Test) which either do not stabilize or, if they do, have unacceptably low task definition.

1. ANALYSIS

Superdiagonal form

Each of the ten tasks to be studied in this report was administered to a group of 18, sometimes 19, volunteer subjects for 15 consecutive working days. The results to be studied take the form of 15 data points for each subject on each task, each point representing that subject's average performance on one day. On one of the tasks studied, Stroop Color-Words, three measures of performance were obtained for each subject on each day; on another task, Arithmetic, two measures of performance were obtained; and on the remaining eight tasks only one measure of performance was obtained. However, one of these tasks, Spoke Trail-Making, existed in two forms, experimental and control; in effect, it constituted two tasks.

Group means and variances for each task on each day have already been presented in another place (Kennedy and Bittner, 1978a). Therefore, while occasional reference may be made in this report to means and variances, the focus will be on correlation. Our concern will be to determine which tasks, if any, are differentially stable. That is, does there come a point in practice on the task where the position of one subject relative to another does not change,

except for random error, as long as external circumstances and subjective conditions remain the same? It should be underscored that instability, as we use the term, does not consist in differential change per se but, rather, in endogenous change, that is, change resulting from practice alone. If the relative ordering among subjects changes in response to some unusual environmental circumstance, for example, an immediately preceding holiday or partial failure of the air conditioning system, the fact is no argument against stability. Indeed, sensitivity to altered environmental or subjective conditions is generally desirable in a performance test. What is not desirable is a change in differential structure as a function of simply taking the test or taking it again.

But how are we to know whether a change in differential structure is endogenous or not? Plainly, this question must have a satisfactory answer or otherwise no attempt to determine task stability can succeed; fortunately it does.

Correlations among trials of practice are usually patterned and, when they are, are always patterned in the same way. This pattern, moreover, is almost always associated with change in the mean or variance; and the more pronounced the change in the mean or variance the rarer it is that this pattern is not found. In addition, this pattern is easily shown to depend on uniform external circumstances. That is, by altering test circumstances or subjective conditions the pattern can be disrupted or even obliterated (Jones, 1969b). Finally, this pattern is naturally and easily explained in terms of continuous endogenous processes, each one taking root at a definite point in practice, continuing for a series of consecutive trials, and then dropping out. On all counts, therefore, this pattern appears to be the correlational counterpart of endogenous differential

change. To determine whether or not a task has stabilized it is sufficient to find out whether or not this pattern is still present.

The pattern in question is "superdiagonal form." The correlations are largest between two trials (in our case, days), one of which immediately follows the other in practice. Correlations between trials that are separated by one trial, for example, days 5 and 7, are smaller. The greater the separation between two trials the smaller the correlation between them is. Hence, the smallest intertrial correlation is found in the upper right-hand corner of the matrix.

The correlations in any one row all involve the same first trial, with the second trial being more and more removed in the practice sequence. Similarly, the correlations in any one column all involve the same second trial, with the first trial coming earlier and earlier in the practice sequence. The requirements of superdiagonal form are that

$$r_{ij} > r_{jk}$$

and

$$r_{ik} < r_{jk} \qquad (i < j < k).$$

That is, the correlations must decrease along the rows to the right and up the columns.

General versus local differential change

The idea of stability does not apply to the task itself or even to all trials of practice on the task but, rather, to all trials past a certain point or, better, to all trials between two points in practice. In our case we will start each analysis by asking whether or not all trials after day 5 are stable, that is, trials 6 through 15. Once, however, we recognize that stability is specific to

a series of trials, not generally beginning with trial 1, then we must also recognize that it exists in two distinct forms.

Consider our own case, that is, trials 6 through 15. If these ten trials are part of an overarching superdiagonal form that begins with trial 1 and continues through trial 15, one consequence is that trial 6 must correlate more strongly with the first five trials than trial 7 does; trial 7 must correlate more strongly with the first five trials than trial 8 does, and so on. In other words, continuing differential change in trials 6 through 15 means, among other things, that each successive trial in this 10-trial series is more and more removed from the first five trials. As practice continues, each successive trial correlates less and less strongly with a fixed set of preceding trials.

At the same time, an overarching superdiagonal form implies that the correlations among trials 6 through 15 have superdiagonal form also. If differential change continues over these ten trials, then intertrial correlation over this same series considered entirely by itself must be patterned in the superdiagonal way.

We will call these two kinds of instability general and local differential change. General differential change takes place relative to an external set of measures. In this report these external measures are always preceding trials of practice on the same task. The idea, however, of general differential change also includes change relative to other tasks. If the correlations between successive trials of practice and a reference test regularly either decrease or increase, the fact is evidence of general differential change. Local differential change is change within a series of consecutive trials. A series of trials that shows no

change with respect to external measures may nevertheless be changing from trial to trial internally. Local change is not just another aspect of a single underlying differential process. Local and general differential change are two different things and do not necessarily occur together.

This last point is central, in part because general and local differential change do not have equal claims on our attention. Suppose that all trials on a task after trial N_o are entirely stable in the general sense. Any local instability that the task may then show is altogether specific to trials after N_o on that task. A structuring of specific variance, however, has very few, if any, practical consequences. It has no appreciable effect on the ability of the trials at issue either to predict external measures or be predicted by them. Let S, consisting of trials s_{m+1}, s_{m+2}, . . ., s_{m+n}, be entirely stable in the general sense and c an outside criterion. Then the correlations between s_{m+i} and c are all the same. Suppose now that the correlations among the s_{m+i} are either all the same or patterned in the superdiagonal way. How much difference does this last variation have on the multiple correlation between S and c? The answer is, not much unless the superdiagonal pattern within S is steep; and steep superdiagonal patterns (hence, rapid local change) do not exist or have not been observed in the absence of general differential change.

In practice, of course, we do not test for all possible kinds of general change. In this report we will look for it only in relation to preceding (or, occasionally, following) trials on the same task, not other tasks or criteria. If, however, a series of trials, S, is stable relative to preceding

trials and the average correlation among trials in S (task definition) does not greatly exceed the average correlation between these trials and just preceding trials, then the probability that the trials in S change appreciably with respect to any external criterion is low. Virtually all of the reliable variance in S is accounted for by its relations with just preceding trials and these trials, by hypothesis, all correlate equally with trials in S. It is technically possible for another task or external measure to show differential change relative to S but not likely and certainly not in a large way.

Local change is, therefore, a distinctly secondary matter. If general stability exists, local change has few, if any, practical consquences. If general stability does not exist, local stability is unlikely and cannot in any case gainsay continuing general change.

In the next two sections we will consider how to test for general and local differential change. We will then take up one or two matters that concern both problems.

Testing for general differential change

Table 1 contains the correlations between the first five and the last ten days on Critical Tracking. Table 2 presents the analysis of variance for these same data, with the row and column effects broken down into linear and nonlinear components.

The row averages show large and overwhelmingly significant increases for the first five days. The regression coefficient for the rows is +0.118 or a predicted increase of +0.472 from day 1 to day 5. This result means that the first five days definitely involve differential change. If we let S consist of the first five trials, then the last ten trials are an external measure; and with

respect to this external measure the trials in S show regular (increasing) change. It is clear, therefore, that Critical Tracking does not stabilize before day 6.

The main result, however, is that the linear component in the column averages is not significant ($F = 0.93$) when tested against the residual mean square. The regression coefficient for the columns is -0.004 per day. Therefore, the predicted average declines by 0.036 from day 6 through day 15.

Note, however, that the nonlinear component in the column averages is significant at the .05 level. Since the hypothesis of endogenous change does not require linear but only monotonic decrease along the columns, the nonlinear component might involve real elements of differential change. Critical Tracking, however, is not a case in point. The nonlinear component is significant because the column averages depart irregularly from the linear regression line, not because the regression line is itself nonlinear. We have already pointed out, however, that irregular variations in the column means do not constitute evidence of endogenous change. On a Monday, for instance, the correlations are sometimes lower than on other days of the week, presumably because the subjects have lost some of their edge and, perhaps, some of their motivation as well over the weekend. The result, though it might well lead to a significant nonlinear column component, does not constitute evidence of endogenous differential change. The change results from an alteration in external circumstances (the weekend), not from practice itself.

We may conclude, therefore, that Critical Tracking is stable with respect to preceding trials after day 5. But is Critical Tracking after day 5 stable with respect to other external measures than preceding trials? The evidence we have on this point is indirect.

The average correlation among days 6 through 15 on Critical Tracking (task definition) is 0.784. The mean correlations, however, between days 4 and 5 and these same ten trials are .760 and .807 respectively. In short, all of the reliable variance in days 6 through 15 is accounted for by their correlations with days 4 and 5. Another external measure, therefore, would almost have to relate to days 6 to 15 through some component that these ten days share with days 4 and 5. It could be, of course, that some components in days 4 and 5 increase over the next ten days and some decrease; but this seems unlikely. If, however, all components that days 4 and 5 share with the next ten days are stable, then the correlations between another external measure and days 6 to 15, mediated as they would almost have to be by one or more of these components, should also be stable.

All in all, therefore, it seems fair to conclude that Critical Tracking is generally stable after day 5.

Testing for local differential change

A. The problem. Table 3 presents the correlations among days 6 through 15 on Critical Tracking. The question is, does this matrix have significant elements of superdiagonal form? That some such elements appear in the matrix is clear from visual inspection. The correlation between days 6 and 15, for example, is smaller (.71) than any correlation in the superdiagonal. In fact, the average of the three correlations in the upper right-hand corner of the matrix (.88 + .71 + .50/3 = .70) is also smaller than any correlation in the superdiagonal. But are these differences significant? That is the question to which we now turn.

B. Background. If a task is stable over a series of trials, S, then except for sampling variations all correlations among trials in S are equal (the matrix is

flat). One possible approach to our question, therefore, might be to test the observed matrix against the hypothesis of a flat matrix at the population level. Fortunately, Lawley has advanced a χ^2 test for precisely this question (Morrison, 1967, pp. 251-252).

Unfortunately, there are serious problems with the use of this test for our purposes. If Lawley's test results in a significant value of χ^2, we will conclude that superdiagonal form is present. That is, the alternative hypothesis to equality of all correlations among trials in S is superdiagonal form. The problem is that Lawley's test may result in a significant value of χ^2 for reasons that have nothing to do with superdiagonal form.

Suppose, for example, that the correlations among trials in S can be perfectly described by a single common factor with unequal factor loadings. Such a matrix will not be flat and, if the loadings are appreciably different, will almost certainly yield a significant result by Lawley's test. This result would be seriously misinterpreted, however, by a conclusion in favor of superdiagonal form. A unit-factor matrix never has superdiagonal form and superdiagonal form can never be explained in terms of a single common factor.

Lawley's test works well enough as long as χ^2 is not significant. That is, if all correlations among trials in S are tenably regarded as equal, then local differential change is absent. If χ^2 is significant, however, we can not conclude that differential change is present. To draw this conclusion we must take some other approach.

A likely possibility is Joreskog's well-known procedures for testing the simplex model (Joreskog, 1970). The simplex is a special kind of superdiagonal form. Joreskog hypothesizes a simplex and then develops, first, a maximum-

likelihood procedure for estimating the theoretical correlations and, then, a χ^2 test for finding out whether or not these correlations are adequate to explain the empirical results.

Unfortunately, Joreskog's test is also inappropriate for our purposes. To make the point directly, suppose that the empirical matrix is essentially flat. Since a flat matrix is a special case of simplicial form, Joreskog's model will fit the data perfectly. Hence, we would conclude in favor of the simplicial interpretation or, more generally, superdiagonal form and differential change. But a flat matrix is the very opposite of differential change. The trouble with the simplex model is that it explains both change and no change. Hence, it cannot distinguish between them.

What we need is a hypothesis, like Lawley's, that posits no differential change and an empirical statistic that reflects it. Then, if we do not obtain significance (and the test we use has sufficient power), we may conclude in favor of stability. On the other hand, if we obtain significance, we can conclude in favor of differential change. In the next section we present such a test.

C. Diagonal comparisons. We begin with a change in notation. Let r_{ij} be the ith correlation, reading down and to the right, in the jth diagonal, reading away from the main diagonal. Thus, the third correlation in the superdiagonal (.92 in table 3) is r_{31}, and the fourth correlation down in the last column (.78 in table 3) is r_{64}. Plainly,

$$1 \leq j \leq n-1$$

and

$$1 \leq i \leq n-j.$$

We advance the model

$$r_{ij} = \delta_j + \epsilon_{ij} ,$$

where δ_j is a fixed effect associated with diagonal j and all ϵ_{ij} are random variables drawn from a normal population with a mean of zero and variance, σ_ϵ^2. Except for error, all correlations in any one diagnonal are, we suppose, equal. The least-squares estimator for δ_j is $\bar{r}_j$; this estimator is unbiased.

The comparisons we propose to make are based on the differences among the diagonals. Let

$$R_j = \sum_{i=1}^{n-k} \sum_{k=j}^{n-1} r_{ik}$$

$$N_j = \sum_{k=j}^{n-1} (n-k) = \frac{(n-j)(n-j+1)}{2}$$

and

$$\bar{R}_j = R_j / N_j .$$

Then consider the quantities

$$C_j = \bar{r}_j - \bar{R}_{j+1} \qquad (j = 1, \ldots, n-2).$$

Each C_j represents the difference between the average correlation in the jth diagonal and the average correlation in all diagonals greater than j, that is, to the "northeast" of the jth diagonal. C_1 is the difference between the average correlation in the superdiagonal and the average of all other correlations in the matrix. C_2 is the difference between the average correlation in the second diagonal and the average of all correlations that span more than three trials. Finally, note that no comparison is defined for j = n-1.

The quantities, C_j, are comparisons; that is, the sum of the coefficients of r_{ij} in C_j vanishes. All correlations in the jth diagonal, (n-j) of them, are

multiplied by $(1/n-j)$; and all correlations in diagonals $(j+1)$ to $(n-1)$, N_{j+1} of them, are multiplied by $(-1/N_{j+1})$. Hence, the sum of all nonzero coefficients in any one C_j is zero.

Furthermore, the $(n-2)$ comparisons, C_j, are all orthogonal to each other. Given any two comparisons, one of them (say, $C_{j'}$) has nonzero coefficients only for correlations all of which are included in $\bar{R}_{j+1}$. Therefore, the sum of cross-products between the coefficients in C_j and $C_{j'}$ is

$$\frac{1}{N_{j+1}}\left[\sum_{i=1}^{n-j'}\frac{1}{n-j'}+\sum_{i=1}^{N_{j'+1}}\left(-\frac{1}{N_{j'+1}}\right)\right]=0.$$

The next step is to calculate the sums of squares attributable to the $(n-2)$ orthogonal comparisons, $SS(C_j)$. Since each comparison has one degree of freedom, $SS(C_j)$ and $MS(C_j)$ are the same. Finally, we will determine the expected value of $MS(C_j)$.

The sum of squares attributable to C_j

$$SS(C_j)=C_j^2\Big/\sum_{i=1}^{n-j}\sum_{j=1}^{n-1}a_{ij}^2,$$

where a_{ij} is the coefficient of r_{ij} in C_j. Hence,

$$\sum_{i=1}^{n-j}\sum_{j=1}^{n-1}a_{ij}^2=(n-j)\left(\frac{1}{n-j}\right)^2+N_{j+1}\left(\frac{-1}{N_{j+1}}\right)^2$$

$$=\frac{(n-j+1)}{(n-j)(n-j-1)}.$$

All told, therefore,

$$SS(C_j)=\frac{(n-j)(n-j-1)}{(n-j+1)}\left(\bar{r}_j-\bar{R}_{j+1}\right)^2.$$

The expected value of this quantity

$$E_{\Delta_j}\left[MS(C_j)\right]=\frac{(n-j)(n-j-1)}{(n-j+1)}\left(\delta_j-\bar{\Delta}_{j+1}\right)^2,$$

where Δ_j is the population counterpart of $\bar{R}_j$. That is,

$$\bar{\Delta}_j=\sum_{i=1}^{n-k}\sum_{k=j}^{n-1}\delta_k\Big/N_j.$$

It only remains to determine a proper estimate of σ_ϵ^2.

In this connection it may be helpful to contrast the analysis into diagonal components with simple analysis of variance. Table 4 presents the sources of variation, sums of squares, and degrees of freedom for an analysis of the n(n-1)/2 correlations into between- and within-diagonal components. This is a simple (one-way) analysis of variance with unequal numbers in the (n-1) groups.

Table 5 presents the sources of variation, sums of squares, and degrees of freedom for the same correlations analyzed by diagonal comparisons. In this light the diagonal comparisons are simply another way of breaking down the between-diagonal variation; in fact, the two sums of squares are equal. That is,

$$\sum_{j=1}^{n-2} SS(C_j) = \sum_{j=1}^{n-1} (n-j)(\bar{r}_{j\cdot} - \bar{r})^2$$

or

$$\sum_{j=1}^{n-2} \frac{(n-j)(n-j-1)}{(n-j+1)} (\bar{r}_j - \bar{R}_{j+1})^2 = \sum_{j=1}^{n-1} (n-j)(\bar{r}_{j\cdot} - \bar{r})^2.$$

Hence, the within-diagonal SS in the simple analysis of variance is the same as the residual SS in the diagonal comparisons. That is,

$$\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} (r_{ij} - \bar{r})^2 - \sum_{j=1}^{n-2} SS(C_j) = \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} (r_{ij} - \bar{r}_{j\cdot})^2.$$

We then find that

$$E[MS_{res}] = E\left[\sum_{i=1}^{n-j} \sum_{j=1}^{n-1} (r_{ij} - \bar{r}_{j\cdot})^2 \Big/ \frac{(n-1)(n-2)}{2}\right]$$

$$= \sigma_\epsilon^2.$$

In both analyses, it should be noted, the degrees of freedom come to one less than the number of correlations in the matrix. In the case of the simple analysis this one degree of freedom is absorbed by μ. In diagonal comparisons it is absorbed by δ_{n-1}.

The residual or within-diagonal SS may be further broken down into linear and nonlinear components as a function of trial or day number. This further analysis, however, is entirely straightforward and will be taken up in the context of a concrete example. The formal framework for an analysis into diagonal comparisons is now in hand.

D. The first five days on Critical Tracking. Table 6 presents the correlations among the first five days on Critical Tracking. Note that the correlations increase strongly in the comparison diagonals. This tendency for the correlations to increase with practice is common where differential change is taking place, especially if it is rapid. It is not, of course, in itself evidence of differential change. A unit factor matrix with increasing factor loadings, for example, would show the same effect. A decision as to whether differential change is present or absent depends solely on the diagonal comparisons. At the same time, regular change within diagonals is an assignable source of variation and should not be included in the error term.

Table 7 presents the analysis into diagonal comparisons for the first five days on Critical Tracking. The diagonal comparisons absorb three degrees of freedom and the linear components within diagonals absorb another three degrees of freedom. The residual term also has three degrees of freedom. None of the F ratios for the diagonal comparisons reaches significance at the .05 level. In this case, however, a conclusion that no local change is taking place is not warranted because with only five trials the analysis does not have sufficient power.

E. The last ten days. Table 8 presents diagonal statistics for the last ten days on Critical Tracking. Included are $\bar{r}_{j}$, $\bar{R}_{j+1}$, C_j, and $SS(C_j)$. The average

correlation, $\bar{r}_j$, is largest in the superdiagonal, second largest in diagonal 2, and smallest in diagonal 8. The rank correlation between $\bar{r}_j$, and j, ρ = -0.79, is significant at the .01 level. This result is sufficient in itself to conclude that some local change is still taking place in the last ten days of Critical Tracking.

The analysis into diagonal comparisons is presented in table 9. None of the diagonal comparisons is significant at the .05 level, although the F ratio for diagonal 1 falls midway between the critical values for the .10 and .05 levels with 1 and 28 degrees of freedom. Certainly, the local tendencies toward superdiagonal form in the last ten trials of Critical Tracking are not sufficient to upset our earlier conclusion that the task stabilizes after day 5.

Transformations and power considerations

The model used in testing for general differential change is

$$\pi_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where μ, α_i, and β_j are fixed effects and ϵ_{ij}, as usual, is a normally distributed random variable with mean equal to zero and variance, σ_ϵ^2. This model implies bilateral compound symmetry. That is, the expected variance along any two rows is the same and the expected variance down any two columns is the same. Similarly the expected covariance between any two rows (or columns) is the same as between any other two rows (or columns).

These consequences may not, of course, be supported by the facts. If the columns evidence stability, it is likely that the column variances will be homogeneous. However, if the rows show differential change, as is usual, it is likely that the later rows with higher average r's will have smaller variances.

Further, if there are changes in the variances along either the rows or columns, it is unlikely that either of the two covariance requirements will be met.

The best approach to this problem is to subject the correlations to Fisher's z transformation. The effect is to "lengthen out" the intervals toward the high end of the correlation scale and this, in turn, tends to homogenize the row and column variances and, hence, to improve the case for compound symmetry in all respects.

A related problem arises in the local analysis. When there is differential change, the correlations within a diagonal tend to increase. This increase, however, may not be linear but negatively accelerated, especially if the correlations exceed .80. The root cause here is the same as in the previous problem, namely, that numerically equal intervals are larger high in the correlation scale than they are lower down. The solution too is the same as before. Fisher's z transformation lengthens out the intervals at the high end of the scale and thereby straightens out the regression with trial or day number. The importance of this straightening out is that it purifies the error term, by removing from it a known source of systematic error.

As intimated earlier in at least two places, power may also be a problem. If we find no significant difference from one column to the next in the general analysis, we conclude that the task has stabilized. Clearly, however, this conclusion raises the power question. What is the probability that we would have obtained a nonsignificant result if the regression coefficient along the columns had been $b_c \neq 0$? With five days or trials, the power of the general analysis is certainly not strong enough to warrant a conclusion of stability. With ten days, however, it is much stronger, although even longer series of trials

would be desirable. There is, however, a limit to the amount of work that empirical investigators can be expected to do in order to meet statistical requirements.

In this connection it is worth noting that failure to meet the requirements of compound symmetry in the general analysis, while it gives the analysis a positive bias, also increases its power.

The power question also arises in relation to the local analysis. If we find no significant diagonal effects, we conclude that superdiagonal form is absent. But what is the probability that we would have obtained a nonsignificant result if ($\delta_d - \overline{\Delta}_{d+1}$) had, in fact, equalled a definite nonzero amount? Here again our only resort is to longer sequences of trials. Finally, power decreases as one moves away from the main diagonal.

2. FIVE STABLE AND WELL-DEFINED TASKS

Code Responses

Table 10 presents the analysis of variance for general change in the last ten days relative to the first five for Code Responses. The key result is the value of F (0.52) for linear change along the columns. The row effects make it clear that stabilization could not be fixed any earlier than the sixth day. The means for the first five days are not only overwhelmingly significant but increase regularly, with one small inversion, from .539 on the first to .781 on the fifth day.

The diagonal averages among the last ten days (table 11) show shallow and certainly innocuous tendencies toward superdiagonal form. The average correlation among the last ten days is .72. This value is definitely low for task

definition, perhaps too low. Certainly, a stable task with good definition would be preferred over Code Responses.

Grammatical Reasoning

The results for Grammatical Reasoning are novel in two respects. The first is that the last ten days are not stable relative to the first five. The linear column component is strongly significant. When this happens, one moves to the next trial and sees if, perhaps, the task may not stabilize from this more advanced point in the practice sequence. In our case, we test the last nine days (days 7 through 15) against the first six. If the linear component is still significant, one takes still another step into the practice sequence and tests again. This process continues until the trials that the subjects have practiced after the one being tested are too few to provide an acceptably powerful test of general differential change. Our convention is to stop at day 10. Thus, we start by testing the last ten days against the first five and end, if no stabilization results, by testing the last five days against the first ten. If the linear component along the columns is still significant, we conclude that as far as our data go the task does not stabilize.

Table 12 presents the average correlations for the last nine days against the first six in Grammatical Reasoning. The linear component along the columns is still significant ($F=20.6$) -- but only because of the low average on day 15. The regression coefficient (average r regressed on day number) is -0.0122. If day 15 is dropped, this same regression coefficient becomes -0.00125; the latter is ten times smaller in absolute terms than the former. Table 13 presents the analysis of variance for general change in days 7 through 14 on Grammatical Reasoning relative to the first six days on the same task. Grammatical Reasoning

is clearly stable over this stretch of eight days. It remains to justify dropping day 15 or, at least, to explain what the implications of doing so are.

The first point we need to recognize is that a task may stabilize for awhile and then start changing again; it may plateau, if you like. Grammatical Reasoning is definitely stable from day 7 through day 14. Its being so, however, in no way requires that it remain stable thereafter. It is perfectly possible that the low column average on day 15 is simply the start of a new phase in differential development on the task. The odds are against it, however.

The 18 volunteers who practiced Grammatical Reasoning knew that day 15 would be their last day on this task. It is possible, therefore, that some of them performed somewhat differently on this last day than they ordinarily did. In other words, the subjects may have responded on day 15 to the fact that this day was to be their last on this task. Such a reaction is an exogenous effect; it is a response to an altered subjective condition (the task is ending). Occurring, however, as it does at the end of practice, the effect of such a reaction is to produce the semblance of linear change.

We cannot be sure, of course, that this interpretation is correct. The main point in its favor is that the regression coefficient up to day 15, that is, from day 7 through day 14, is essentially flat. In addition, nonlinear change over this series of eight trials, while significant, is modest. In general, the column averages are not bouncing around a great deal. Hence, the marked fall on day 15 requires more of an explanation. Finally, the effect is not isolated; we will see it again in other tasks. We conclude, therefore, that Grammatical Reasoning would probably remain stable more or less indefinitely after

day 6, provided subjective conditions also remained the same.

Table 14 presents the diagonal averages for days 7 through 14 on Grammatical Reasoning. With the exception of diagonal 6 the fall-off away from the main diagonal is regular. It is also quite shallow, however, and poses no problems of any consequence. Task definition from days 7 through 14 is high, 0.881.

Arithmetic

Arithmetic also presents novel problems. This task has two measures, number attempted and number correct. The results for general change in the last ten days relative to the first five are presented in tables 15 and 16. Note that linear change along the columns is significant in both cases. The regression coefficients, however, are both positive! They are also very small and, as it happens, equal, +0.0027. What are we to make of this state of affairs?

The intertrial correlations for Arithmetic are very high and tightly bunched. The correlations on number attempted range from .85 to .97 and on number correct from .86 to .97. The residual term is miniscule. The change along the columns for all 10 days comes to only +0.024, certainly a small amount. But what about its direction? How are we to account for the increase in column correlations over the last ten days.

The answer lies in the row effects. The changes here are much larger and more significant than the column trends. And they too go the wrong way! That is, the row means tend to decrease from day 1 through day 5. On number attempted, for example, the average correlation on day 1 is .952 and on day 5 it is .895.

These results altogether exclude differential change of the superdiagonal sort. In the sense that we have been using the term, Arithmetic is stable from

day 1. For the first week performance seems to acquire rather more random elements and thereafter gradually to lose them. The effect is to create a shallow bow in the correlation pattern--but a bow created by slow swings in specific variance. The correlation matrix is a Spearman unit hierarchy with somewhat smaller factor loadings around day 5 than either before or after. This pattern, however, is consistent with stability, in fact, stability of a rock-solid order.

When they came into service, our volunteers had already learned all the arithmetic they would ever know. They were already stabilized on this task and no change would subsequently occur in common variance, that is, between one day and any other. Local change in Arithmetic is negligible and task definition very high, 0.949 for number attempted and 0.948 for number correct.

The Stroop Test

The Stroop Test yields three measures: blocks/words, colored words, and colored blocks. Tables 17 and 18 present the analyses for general change in the last ten days for the first two measures. Both measures are stable after day 5.

Colored blocks presents a more complicated picture. The F ratio for the linear column component is 7.47, just short of significance at the .01 level. The regression coefficient for the column averages is -0.005 or a decrease of 0.045 over the 10-day period. As in Grammatical Reasoning, however, this decrease stems entirely from a low average for day 15. If day 15 is dropped, the coefficient becomes ever so slightly positive, +0.0005; and linear column change is no longer significant (F=0.33).

Our interpretation of these results is the same as for Grammatical Reasoning, that is, that in all probability the low average on day 15 is attributable to an

altered subjective condition (the task is ending). In this case the interpretation is supported by the clear stability of the other two measures on the same task. We conclude, therefore, that the Stroop Test is stable on all three measures after day 5.

Table 19 presents the diagonal averages for the last ten days on colored blocks. Superdiagonal form appears to be entirely absent. The same, more or less, is true of blocks/words and colored words. Task definition is good for all three measures, 0.827 for blocks/words, 0.867 for colored words, and 0.883 for colored blocks.

Two-Dimensional Tracking

Although it was not discovered until after the experiment was completed, the software used in Two-Dimensional Tracking contained a "dead" spot. When the cursor was placed on this spot, it remained there with no further control movements. When the experimenters finally discovered this dead spot, they interviewed the subjects concerning it. Several of the subjects reported that they discovered the spot around day 8 and subsequently made use of it from time to time to "beat" the task. The existence of this spot and its discovery by some subjects (not all, apparently) fundamentally altered the task after day 8 or thereabouts.

Table 20 presents the analysis of variance for days 6-9 on Two-Dimensional Tracking relative to days 1-5 on the same task. The linear column component is vanishingly small. The regression coefficient along the columns is -0.0002. With only four trials as a basis, the test for linear column change is admittedly not powerful. Nevertheless, our judgment is that Two-Dimensional Tracking stabilizes after five days. Note, by the way, the strong linear component down the rows. The row means show regular change through day 5. Hence, stabilization cannot be said to begin any earlier than day 6.

Table 21 presents the correlations between days 6-9 (the so-regarded stabilized trials) and days 10-15. For the first three days (10-12) the correlations hold up well, but then they fall dramatically. It appears, therefore, that beginning sometime in the third week Two-Dimensional Tracking underwent rapid differential change, presumably because of the dead spot and its discovery. It is bothersome, of course, that the dead spot seems not to have had an immediate effect on differential processes. On the other hand, it appears to have had no effect at all on group processes. The mean and variance show no disturbances at all as a result of the dead spot. Looking at the two curves, no one would suspect that anything unusual had happened toward the end of the second week or at any other time in practice. This point has obvious methodological importance since it underscores the sensitivity of differential processes to changes that would otherwise go unnoticed.

Task definition in days 6-9 on Two-Dimensional Tracking is passable but not good, 0.767.

3. FOUR UNSTABLE OR ILL-DEFINED TASKS

Complex Counting

Table 22 presents the analysis of variance for general change in the last ten days of Complex Counting relative to the first five days on the same task. The F ratio for the linear column component is strongly significant (F=2.63). The linear column component is still significant when the last nine days are tested for general change relative to the first six days. In fact, the linear column component remains significant for the last eight, seven, six and five days. Table 23 presents the results for days 11-15 versus the first ten days.

The F ratio for the linear column component is smaller than in table 22 but still strongly significant (F=14.1).

This time, moreover, the decrease along the columns is not due solely to the last day. The regression coefficient for days 11-15 is -0.021 and for days 11-14 it is -0.017, smaller, to be sure, but not greatly so. The linear column component for days 11-14 relative to the first ten days is, moreover, still significant, albeit at the .05 level.

These gradually lowering levels of significance as one moves further and further into the practice sequence suggest that at some point Complex Counting does, indeed, stabilize. That point, however, is not reached after ten days of practice.

Time Estimation

The results for Time Estimation are much the same as for Complex Counting. Table 24 presents the average correlations of the last five days with each of the first ten days. The main point is that these averages increase right up to and including day 10. Hence, the first ten trials are changing relative to the last five as an external measure. There is no possibility, therefore, that Time Estimation stabilizes any earlier than day 11. The question is, does it stabilize then?

Table 25 presents the analysis of variance for general change in the last five days relative to the first ten days. Linear change from row to row is enormous; it yields the largest F value we have seen, thus confirming the conclusion already reached that differential change continues through day 10. Since linear change from column to column is also strongly significant, F=15.4, it would seem that Time Estimation is still not stable after ten days.

There is, however, reason to pause for a moment. The correlation average for day 15 is lower than for any other day in the last week. If day 15 is omitted, the regression coefficient along the columns drops from -0.038 to -0.018 and the linear column component is no longer significant. If the low $\bar{r}$ for day 15 can be reasonably attributed to altered subjective conditions (the task is ending), then a case could be made that Time Estimation stabilizes after ten days. We do not think, however, that the low $\bar{r}$ on day 15 can be so attributed.

In the first place, both column and row averages bounce around a good deal in Time Estimation. The drop, for example, from day 12 to day 13 is almost as large as the drop from day 14 to day 15. Similarly, as can be seen in table 22, the row averages also change sharply from one day to the next. Hence, the drop from day 14 to day 15 is by no manner of means a unique occurrence in this set of data.

In the second place, local change appears to continue in Time Estimation through the last five days. Table 26 presents the diagonal statistics for days 11 through 15 and table 27 the analysis into diagonal components. The F ratio for diagonal 3 (7.1) is significant at the .10 level. In addition, all three comparisons are positive, and the probability of this result is 0.125 by itself. This last consideration, that is, how many of the C_j are positive and how many negative, tests the same hypothesis as do the diagonal components, namely, that local differential change continues in the last five days of practice; moreover, it does so independently. Combining these four tests (Winer, 1967, pp. 49-50) yields a value of χ^2 which is significant at the .07 level. Local change, therefore, appears to be at least likely in days 11 through 15.

On these grounds we conclude that Time Estimation does not stabilize after 10 days and probably not after 15 days. Task definition in the last five days is marginal, 0.718.

Letter Search

Task definition for Letter Search is unacceptably low. The average correlation among days 6-15 is 0.422 and among days 11-15 only a little better, 0.510. Whatever else may be said about it, Letter Search is not a suitable task for performance testing, whether it stabilizes or not. Hence, we pursue its analysis no further.

Spoke Trail-Making

Spoke Trail-Making was, in effect, two tasks, a standard or control task and an experimental variation. On the evidence in hand neither of these tasks meets the requirements for performance testing. Task definition for the experimental form is too low. The average correlation among days 6-15 and 11-15 are 0.444 and 0.502 respectively.

The case of the control task is more complicated. Testing the last ten days against the first five one finds an F ratio for the linear column component which is significant at the .05 level. Testing the last five days against the first ten one finds a larger F ratio for the linear column component, significant at the .01 level. These results, however, are due entirely to a low $\bar{r}$ on the next to last day, day 14. If we exclude day 14, the average r's of days 6-15 with the first five days range from a low of .724 to a high of .866. The average correlation of day 14 with the first five days is 0.452! It may be that this low $\bar{r}$ was due to some unrecorded change in test circumstances or subjective conditions. If so, then the standard form of Spoke Trail-Making is stable and has good task definition. On the existing evidence, however, we have no grounds for excluding day 14. We can

hardly argue that the subjects "wound down" on the day before the test ended but not the last day. For the time being, therefore, we regard the control form of Spoke Trail-Making as unstable.

REFERENCES

Alvares, K. M. & Hulin, C. L. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 1972, 14, 295-308.

Alvares, K. M. & Hulin, C. L. An experimental evaluation of a temporal decay in the prediction of performance. Organizational Behavior and Human Performance, 1973, 9, 169-185.

Dunham, R. B. Ability-skill relationships: An empirical explanation of change over time. Organizational Behavior and Human Performance, 1974, 12, 372-382.

Jones, M. B. Differential processes in acquisition. In E. A. Bilodeau and I. McD. Bilodeau (Eds.), Principles of skill acquisition. New York: Academic Press, 1969. (a)

Jones, M. B. Knowledge of results and intertrial correlations in a simple motor task. Journal of Motor Behavior, 1969, 1, 331-340. (b)

Jones, M. B. A two-process theory of individual differences in motor learning. Psychological Review, 1970, 77, 353-360. (a)

Jones, M. B. Rate and terminal process in skill acquisition. American Journal of Psychology, 1970, 83, 222-236. (b)

Jones, M. B. Individual differences. In R.N. Singer (Ed.), The psychomotor domain. Philadelphia: Lea and Febiger, 1972.

Joreskog, K. G. Estimation and testing of simplex models. British Journal of Mathematical and Statistical Psychology, 1970, 23, 121-145.

Kennedy, R. S., & Bittner, Jr., A. C. The development of a Performance Evaluation Test for Environmental Research (PETER). In Productivity enhancement in Navy systems. San Diego, California: Naval Personnel Research and Development Center, October, 1977.

Kennedy, R. S., & Bittner, Jr., A. C. The stability of complex human performance for extended periods: Applications for studies of environmental stress. Presented at the Aerospace Medical Association, New Orleans, Lousiana, May, 1978 (printed in Preprints.). (a)

Kennedy, R. S., & Bittner, Jr., A. C. Progress in the analysis of a Performance Evaluation Test for Environmental Research (PETER). Proceedings of the 22nd Annual Meeting of the Human Factors Society, Santa Monica, California, 1978. (b)

Morrison, D. F. Multivariate statistical methods. New York: McGraw-Hill, 1967.

Winer, B. J. Statistical principles in experimental design (Second Edition). New York: McGraw-Hill, 1971.

Table 1

Correlations between the first five and the last ten days of practice on Critical Tracking in 18 Navy volunteers.

First Five Days	6	7	8	9	10	11	12	13	14	15	$\bar{X}$
1	.47	.58	.40	.38	.05	.17	.25	.24	.34	.25	.313
2	.63	.52	.59	.57	.38	.46	.49	.57	.64	.79	.564
3	.73	.62	.80	.74	.67	.76	.84	.64	.74	.73	.727
4	.85	.66	.85	.81	.62	.74	.82	.76	.76	.73	.760
5	.89	.73	.89	.82	.68	.77	.86	.78	.85	.80	.807
$\bar{X}$	.714	.622	.706	.664	.480	.580	.652	.598	.666	.660	.634

Table 2

Analysis of variance for general change in the last ten days of Critical Tracking relative to the first five days on the same task.

Source	SS	df	MS	F
Rows	1.624	4	0.406	58.0 *
linear	1.402	1	1.402	200.3 *
nonlinear	0.222	3	0.074	10.6 *
Columns	0.213	9	0.024	3.4 *
linear	0.007	1	0.007	0.9
nonlinear	0.206	8	0.026	3.7 *
Residual	0.251	36	0.007	
Total	2.087	49		

*Significant at the .01 level.

Table 3

Correlations among days 6 through 15 on Critical Tracking

Day	6	7	8	9	10	11	12	13	14	15
6	—	.81	.88	.83	.68	.77	.80	.76	.88	.71
7		—	.72	.65	.48	.55	.56	.47	.72	.50
8			—	.92	.87	.88	.90	.82	.89	.79
9				—	.83	.89	.90	.81	.90	.78
10					—	.92	.87	.79	.82	.74
11						—	.90	.72	.83	.69
12							—	.82	.85	.77
13								—	.91	.86
14									—	.86
15										—

Table 4

Sources of variation, sums of squares, and degrees freedom for the analysis of a correlation matrix into between and within diagonal components.

Source	SS	Degrees of freedom	
		n	n=10
Between diagonals	$\sum_{j=1}^{n-1} (n-j)(\bar{r}_j - \bar{r})^2$	$(n-2)$	8
Within diagonals	$\sum_{i=1}^{n-2} \sum_{j=1}^{n-1} (r_{i,j} - \bar{r}_j)^2$	$\frac{(n-1)(n-2)}{2}$	36
Total	$\sum_{i=1}^{n-2} \sum_{j=1}^{n-1} (r_{i,j} - \bar{r})^2$	$\frac{n(n-1)}{2} - 1$	44

Table 5

Sources of variation, sums of squares, and degrees of freedom for the analysis of a correlation matrix into diagonal-comparison and residual components.

Source	SS	Degrees of freedom: n	Degrees of freedom: n=10
Diagonal comparisons	$\sum_{j=1}^{n-2} \frac{(n-j)(n-j-1)}{(n-j+1)}\left(\bar{r}_{j} - \bar{R}_{j+1}\right)^2$	$(n-2)$	8
Residual	$\sum_{i=1}^{n-1}\sum_{j=1}^{n-i}\left(r_{ij} - \bar{r}\right)^2 - \sum_{j=1}^{n-2} SS(C_j)\frac{(n-1)(n-2)}{2}$	36	
Total	$\sum_{i=1}^{n-1}\sum_{j=1}^{n-i}\left(r_{ij} - \bar{r}\right)^2$	$\frac{n(n-1)}{2} - 1$	44

Table 6

Correlations among the first five days on Critical Tracking In 18 volunteer subjects.

Day	1	2	3	4	5
1	---	.29	.47	.49	.49
2		---	.62	.69	.75
3			---	.79	.87
4				---	.93
5					---

Table 7

Analysis by diagonal comparisons for the correlations among the first five days on Critical Tracking.

Source of Variation	SS	df	MS	F
Diagonal comparison	0.0286	3	0.0095	2.79
diagonal 1	0.0023	1	0.0023	0.68
diagonal 2	0.0150	1	0.0150	4.41
diagonal 3	0.0113	1	0.0113	3.32
Linear trend	0.3322	3	0.1107	32.56
diagonal 1	0.2184	1	0.2184	64.24
diagonal 2	0.0800	1	0.0800	23.53
diagonal 3	0.0338	1	0.0338	9.94
Residual	0.0101	3	0.0034	...
Total	0.3709	9		

Table 8

Diagonal statistics for the last four days on Critical Tracking.

Diagonal (j)	$(n-j)$	$\bar{r}_j$	R_{j+1}	C_j	$SS(C_j)$
1	9	0.854	0.767	0.087	0.0545
2	8	0.824	0.751	0.073	0.0331
3	7	0.783	0.740	0.043	0.0097
4	6	0.742	0.739	0.003	0.0000
5	5	0.758	0.730	0.028	0.0026
6	4	0.735	0.727	0.008	0.0002
7	3	0.757	0.697	0.060	0.0054
8	2	0.690	0.710	-0.020	0.0003

Table 9

Analysis by diagonal comparisons for the last ten days on Critical Tracking.

Source of Variation	SS	df	MS	F
Diagonal comparisons	0.1058	8	0.0132	0.88
diagonal 1	0.0545	1	0.0545	3.63
diagonal 2	0.0331	1	0.0331	2.21
diagonal 3	0.0097	1	0.0097	0.65
diagonal 4	0.0000	1	0.0000	0.00
diagonal 5	0.0026	1	0.0026	0.17
diagonal 6	0.0002	1	0.0002	0.01
diagonal 7	0.0054	1	0.0054	0.36
diagonal 8	0.0003	1	0.0003	0.02
Linear trend	0.1086	8	0.0136	0.91
diagonal 1	0.0068	1	0.0068	0.45
diagonal 2	0.0001	1	0.0001	0.01
diagonal 3	0.0066	1	0.0066	0.44
diagonal 4	0.0082	1	0.0082	0.55
diagonal 5	0.0078	1	0.0078	0.52
diagonal 6	0.0065	1	0.0065	0.43
diagonal 7	0.0004	1	0.0004	0.03
diagonal 8	0.0722	1	0.0722	4.83
Residual	0.4199	28	0.0150	____
Total	0.6343	44	____	____

Table 10

Analysis of variance for general change in the last ten days of Code Responses relative to the first five days on the same task.

Source	SS	df	MS	F
Rows	0.380	4	0.095	5.94
linear	0.312	1	0.312	19.50
nonlinear	0.068	3	0.023	1.44
Columns	0.133	9	0.015	0.92
linear	0.008	1	0.008	0.50
nonlinear	0.125	8	0.016	1.00
Residual	0.563	36	0.016	
Total	1.076	49		

Table 11

Diagonal averages for the last ten days on Code Responses.

Diagonal (j)	(n-j)	$\bar{r}_j$
1	9	0.746
2	8	0.769
3	7	0.703
4	6	0.685
5	5	0.698
6	4	0.675
7	3	0.693
8	2	0.755

Table 12

Average correlations between days 7 through 15 and days 1 through 6 on Grammatical Reasoning, organized by the former.

Last nine days	Average correlation with first six days
7	0.708
8	0.735
9	0.767
10	0.740
11	0.678
12	0.708
13	0.668
14	0.775
15	0.552

Table 13

Analysis of variance for general change in days 7 through 14 on Grammatical Reasoning relative to the first six days on the same task.

Source	SS	df	MS	F
Rows	0.6484	5	0.1297	68.3*
linear	0.5876	1	0.5876	309.3*
nonlinear	0.0608	4	0.0152	8.0*
Column	0.0634	7	0.0091	4.8*
linear	0.0004	1	0.0004	0.2
nonlinear	0.0630	6	0.0105	5.5*
Residual	0.0649	35	0.0019	
Total	0.7767	47		

*Significant at the .01 level.

Table 14

Diagonal averages for days 7 through 14 on Grammatical Reasoning.

Diagonal (j)	(n-j)	$\bar{r}_j$
1	7	0.889
2	6	0.878
3	5	0.874
4	4	0.872
5	3	0.860
6	2	0.910

Table 15

Analysis of variance for general change in the last ten days of Arithmetic relative to the first five days on the same task. The measure is "number attempted."

Source	SS	df	MS	F
Rows	0.0176	4	0.0044	9.8*
linear	0.0137	1	0.0137	30.4*
nonlinear	0.0039	3	0.0013	2.9**
Columns	0.0066	9	0.0010	1.4
linear	0.0030	1	0.0030	6.7**
nonlinear	0.0056	8	0.0007	1.6
Residual	0.0161	36	0.00045	
Total	0.0423	49	---	

* Significant at the .01 level.

** Significant at the .05 level.

Table 16

Analysis of variance for general change in the last ten days of Arithmetic relative to the first five days on the same task. The measure is "number correct."

Source	SS	df	MS	F
Rows	0.0154	4	0.0038	10.0*
linear	0.0110	1	0.0110	28.5*
nonlinear	0.0044	3	0.0015	3.8**
Column	0.0102	9	0.0011	2.8**
linear	0.0029	1	0.0029	7.5*
nonlinear	0.0073	8	0.0009	2.4
Residual	0.0139	36	0.0004	---
Total	0.0395	49		

* Significant at the .01 level.

** Significant at the .05 level.

Table 17

Analysis of variance for general change in the last ten days of the Stroop test relative to the first five days on the same task. The measure is "blocks/words."

Source	SS	df	MS	F
Rows	0.5252	4	0.1313	50.5*
linear	0.4225	1	0.4225	162.5*
nonlinear	0.1027	3	0.0342	13.2*
Columns	0.0703	9	0.0078	3.0*
linear	0.0052	1	0.0052	2.0
nonlinear	0.0651	8	0.0081	3.1*
Residual	0.0919	36	0.0026	
Total	0.6874	49		

*Significant at the .01 level.

Table 18

Analysis of variance for general change in the last ten days of the Stroop Test relative to the first five days on the same task. The measure is "colored words."

Source	SS	df	MS	F
Rows	0.7037	4	0.1759	83.8*
linear	0.5227	1	0.5227	248.9*
nonlinear	0.1810	3	0.0603	28.7
Columns	0.1360	9	0.0151	7.2*
linear	0.0000	1	0.0000	0.0
nonlinear	0.1360	8	0.0170	9.0*
Residual	0.0767	36	0.0021	
Total	0.9163	49		

*Significant at the .01 level.

Table 19

Diagonal averages for the last ten days on the Stroop test, with "colored blocks" as the measure.

Diagonal (j)	(n-j)	$\bar{r}_j$
1	9	0.886
2	8	0.881
3	7	0.871
4	6	0.907
5	5	0.882
6	4	0.883
7	3	0.890
8	2	0.845

Table 20

Analysis of variance for days 6-9 on Two-Dimensional Tracking relative to days 1-5 on the same task.

Source	SS	df	MS	F
Rows	0.2262	4	0.0566	25.7*
linear	0.1836	1	0.1836	83.5*
nonlinear	0.0426	3	0.0142	6.5*
Columns	0.0042	3	0.0014	0.6
linear	0.0000	1	0.0000	0.0
nonlinear	0.0042	2	0.0021	1.0
Residual	0.0267	12	0.0022	----
Total	0.2572	19		

*Significant at the .01 level.

Table 21

Correlations between days 6-9 and days 10-15 on Two-Dimensional Tracking.

Day	10	11	12	13	14	15	$\bar{r}$
6	.71	.39	0.64	.38	.21	.21	0.423
7	.84	.62	.77	.46	.30	.42	0.568
8	.77	.52	.71	.34	.07	.27	0.447
9	.70	.77	.78	.37	.12	.34	0.513
$\bar{r}$	0.755	0.575	0.725	0.388	0.175	0.310	0.488

Table 22

Analysis of variance for general change in the last ten days of Complex Counting relative to the first five days on the same task.

Source	SS	df	MS	F
Rows	0.3411	4	0.0853	27.5*
linear	0.2948	1	0.2948	95.1*
nonlinear	0.0463	3	0.0154	5.0*
Columns	0.1277	9	0.0142	4.6*
linear	0.0814	1	0.0814	26.3*
nonlinear	0.0463	8	0.0058	1.9
Residual	0.1121	36	0.0031	
Total	0.5809	45		

*Significant at the .01 level.

Table 23

Analysis of variance for general change in the last five days of Complex Counting relative to the first ten days on the same task.

Source	SS	df	MS	F
Rows	0.2827	9	0.0314	10.1*
linear	0.0682	1	0.0682	22.0*
nonlinear	0.2145	8	0.0268	8.6*
Columns	0.0515	4	0.0219	4.2*
linear	0.0437	1	0.0437	14.1*
nonlinear	0.0078	3	0.0026	0.8
Residual	0.1118	36	0.0031	
Total	0.4460	45		

*Significant at the .01 level.

Table 24

Average correlations of the first ten days on Time Estimation with the last five days on the same task.

Day	Average correlation with last five days
1	-0.180
2	+0.068
3	-0.138
4	-0.034
5	+0.232
6	+0.522
7	+0.454
8	+0.522
9	+0.474
10	+0.740

Table 25

Analysis of variance for general change in the last five days of Time Estimation relative to the first ten days on the same task.

Source	SS	df	MS	F
Rows	4.6342	9	0.5149	54.8 *
linear	3.9646	1	3.9646	421.8 *
nonlinear	0.6696	8	0.0837	8.9 *
Columns	0.2821	4	0.0705	7.5 *
linear	0.1444	1	0.1444	15.4 *
nonlinear	0.1377	3	0.0459	4.9 *
Residual	0.3373	36	0.0094	
Total	5.2536	45		

* Significant at the .01 level.

Table 26

Diagonal statistics for the last five days in Time Estimation.

Diagonal (j)	(n-j)	$\bar{r}_j$	$\bar{R}_{j+1}$	C_j	$SS(C_j)$
1	4	0.790	0.670	0.120	0.0346
2	3	0.733	0.607	0.126	0.0238
3	2	0.740	0.340	0.400	0.1067

Table 27

Analysis by diagonal comparisons for the last five days of Time Estimation.

Source of Variation	SS	df	MS	F
Diagonal comparisons	0.1651	3	0.0550	3.7
diagonal 1	0.0346	1	0.0346	2.3
diagonal 2	0.0238	1	0.0238	1.6
diagonal 3	0.1067	1	0.1067	7.1
Linear trend	0.2254	3	0.0751	5.0
diagonal 1	0.0372	1	0.0372	2.5
diagonal 2	0.1682	1	0.1682	11.2
diagonal 3	0.0200	1	0.0200	1.3
Residual	0.0451	3	0.0150	
Total	0.4356	9		